Localization and Characterization of Mouse-Human Alignments Within the Human Genome. Does Evolutionary Conservation Suggest Functional Importance?

Sunil K. Saluja MD[1,2], Isaac Kohane MD[2,4]
[1]Children's Hospital Informatics Program, Boston, MA, [2]Harvard Medical School

Abstract

In an attempt to validate the use of evolutionary conservation as a method to identify putative regulatory elements, we have quantified the frequency of Single Nucleotide Polymorphisms (SNPs) within the most tightly conserved regions across the entire Human Genome. Our results show that conserved non-coding sequences have a significantly lower SNP frequency than their exonic counterparts, which suggests that these regions are functionally important.

Background

Functional annotation of the human genome has been a major challenge in bioinformatics. While early efforts were aimed at identifying exons, complete annotation must also include functionally important non-protein-coding sequences. Alterations in these sequences may influence gene transcriptional control, affecting both disease and health. Identifying these regions, however, is a difficult task. Non-transcribed sequence accounts for greater than 90% of the human genome. Empirically, regulatory regions in mammalian genomes are not at distinct positions relative to genes, but can be present anywhere along the genome.

One approach to reducing the search space is through the identification of evolutionarily conserved genomic segments.[1] This approach may be validated through the identification of SNPs within these regions. If conserved regions are functionally important, they should have a relatively low frequency of identified SNPs.

Methods

Mouse-human alignment maps were obtained from the UCSC Human Genome Project / Mouse genome sequence Consortium, as the source of conservation information. [2] The most tightly conserved data set, covering the top 6th percentile of conserved regions across the entire Human Genome was used to query SNPper, a Single Nucleotide Polymorphism retrieval system, based on dbSNP and SNP Consortium databases.[3] Regions were also localized relative to identified genes, using NCBI Reference Sequence annotation of gene location. Conserved regions were classified as exonic, intronic, intergenic, flanking, and untranslated. Flanking regions spanned across an exon boundary. Un-translated regions were transcribed, but not translated.

Results

Characterization of Conserved Sequences

Conserved Sequence	**Exonic**	**Flanking**	**Intronic**	**Intergenic**
number	18,165	11,558	109,843	641,700
percent	2.3	1.4	16.5	79.8
Avg Size(bp)	348	450	188	201
SNP frequency	0.00407	0.00082	0.00057	0.00062

(Table 1)

The table above characterizes all tightly conserved human-mouse alignments across the Human Genome Nov2002 release. A majority of these conserved regions were non-protein coding. More conserved non-protein coding regions were found between genes, than within introns. The frequency of known SNPs was highest in conserved exonic regions. As a subset of exonic regions, untranslated regions had a lower SNP frequency of 0.00088.

Discussion

These results suggest an important functional role for conserved non-protein coding regions within the genome. This is supported by a lower SNP frequency in these regions and the preponderance of these segments within the set of all tightly conserved regions. Some of these differences may also be attributed to potentially synonymous SNPs within exons, while true functional elements may require stricter adherence to the primary sequence in order to retain function. While these regions have been filtered with a methodology designed to reduce the noise generated by repetitive elements, further filtering at this point may help generate a smaller list of putative regulatory elements.

References

[1] Dermitzakis ET, Clark AG. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol*. Jul:19(7) 2002

[2] Kent W, Haussler D, The Human Genome Browser at UCSC *Genome Research*, May:12(5) 2002.

[3] Riva A, snpper.chip.org, 2001